International Journal of Research in Computer Science
eISSN 2249-8265 Volume 2 Issue 4 (2012) pp. 1-5

www.ijorcs.org

A NOVEL STATISTICAL METHOD FOR THERMOSTABLE PROTEINS DISCRIMINATION

Elham Nikookar[1], Kambiz Badie[2], Mehdi Sadeghi[3]

[1]*Department of Algorithm and Calculations, Faculty of Engineering, University of Tehran, Tehran, Iran*
Email: e.nikookar@ut.ac.ir

[2]*Research Institute for ICT, Tehran, Iran*
Email: k_badie@itrc.ac.ir

[3]*National Institute of Genetic Engineering and Biotechnology, Tehran, Iran*
Email: sadeghi@nigeb.ac.ir

Abstract: *In this study, we used features that can be extracted from protein sequences to discriminate mesophilic, thermophilic and hyper-thermophilic proteins. Amino acid frequency, dipeptide amino acid frequency and physical-chemical features are used in this study. The effect of mentioned features on proposed discrimination algorithm was evaluated both separately and in combination. Statistical methods are used in the proposed algorithm. The results of implementing the algorithm on a dataset containing 239 mesophilic proteins, 69 thermophilic proteins and 59 hyper-thermophilic proteins show the effect of each bunch of features on the evaluation measures.*

Keywords: *protein mesophile thermophile, discrimination, thermostability, amino acid frequency, feature extraction.*

I. INTRODUCTION

Proteins are one of the four major classes of biological macromolecules. They are linear chains of typically ~50-1200 amino acids. Some of the proteins functions include catalyzing metabolic reactions, chemical signal transduction, and forming the physical skeleton of some cellular components. Most types of protein molecule fold into a well-defined shape under physiological conditions and this shape is uniquely determined by the amino acid sequence of the protein. The shape of the molecule facilitates its biochemical function [1]. There are twenty amino acid types that occur in living things [2, 3]. Amino acids are sometimes called residues when covalently bonded together to form a protein molecule.

Organisms that thrive in very high temperatures have been actively studied since the discovery of Thermophiles aquatics in the hot springs of Yellowstone in the 1960's [4]. Since then, thermostable proteins, because of their overall inherent stability, have become a part of a number of commercial applications. If we know the protein thermostability very well, it would be helpful to know better the folding mechanism and the function of protein [5]. To understand the principles that rule protein thermostability has become of great interest in basic research as well as in industrial applications. Several investigations [6, 7] where carried out and researchers found that the protein amino acid composition was correlated to its thermostability.

Initial research conducted in the field of protein thermal stability dates back to three decades ago [2]. The reason is that thermal instability of proteins at high temperature is a barrier against the functional development of proteins. Haney [8] summarized the net change in amino acid composition between mesophilic and thermophilic proteins. The thermophilic proteins were characteristically reduced in Ser, Asn, Gln, Thr and Met, and were increased in Ile, Arg, Glu, Lys and Pro. The comparison of residue contents in hyper-thermophilic and mesophilic proteins on the basics of the genome sequences of eight mesophilic and seven hyper-thermophilic organisms showed that more charged residues existed in hyper-thermophilic proteins than in mesophilic proteins [9].

However, it was difficult to find the influence of dipeptide composition on protein thermostability [10] from the diverse collection of studies. Proteins that have similar amino acid composition vary in dipeptide composition; while, amino acids function with the help of other residues nearby in sequence or space. Accordingly, the function of specific amino acid was influenced by its neighboring amino acid in sequence or space, while the dipeptide reflects the influence in sequence. Hence, the dipeptide composition may be correlated to protein thermostability. For the past two decades, the methods based on dipeptide composition have been used for predicating protein structure class [11, 12]. Structurally, proteins are in different 3-D dimensional shapes which their shape determines biological and chemical duty of protein [1]. Shape of a protein, has a one-to-one relationship with amino acid sequence of it.

If we find effective factors in thermal stability of proteins, using reasonable design, more stable proteins can be achieved. Previous investigations [13] show that proteins thermal stability depends on its amino acid composition. In previous studies, classification is done only based on thermophile and mesophile classes [5]. This division leads to limitation of methods on only some applications and specific datasets. In this paper, in addition to amino acids frequencies, we used other features that can improve proteins discrimination. Also, proteins are divided into mesophile, thermophile and hyper-thermophile classes. This makes possible to generalize proposed method to more general data sets and applications.

II. DATASET

Scientists divide organisms based on their thermal stability into two classes: thermophilic proteins that have an affinity for elevated temperatures and grow in temperatures between 45-80 degrees Celsius and hyper-thermophilic proteins that their optimum growth temperature is over 80 degrees Celsius. Mesophilic proteins function in 15-45 degrees Celsius. Hyper-mesophilic proteins have optimum growth temperature below 15 degrees Celsius [13].

In this paper, we used the dataset of [14], which contains 58 hyper-thermophilic proteins and 118 mesophilic proteins homolog with them and also 69 thermophilic proteins and 121 mesophilic proteins homolog with them. All the proteins have been downloaded from PDB (Protein Data Bank) [15] repository. Structure of each sample of dataset contains the amino acid sequence with the length of 50-1200 from which we extracted elements of feature vector [13].

Table 1: Chemical and physical properties

ID	Property	AA	ID	Property	AA	ID	Property	AA
1	aromatic	HFWY	17	carboxyl	DE	33	Polar/hydrophobilic	RNDEQHKSTWY
2	UV absorbance	FWY	18	carbonyl	NDEQ	34	hydrophobic	ACILMFPWYV
3	Single aromatic ring	FY	19	imidazol	H	35	Very hydrophobic	ACILMFV
4	heteroaromatic	HW	20	guanidio	R	36	Weak hydrophobic	PWY
5	aliphatic	GAILVP	21	amino	RK	37	H-bonding	RNDCEQHKSTWY
6	branched	ILTV	22	Systematical alpha-C	G	38	H-acceptor	NDEQHSTY
7	Branched beta-carbon	ITV	23	alkyl	AILV	39	H-donor	RNCQHKSTWY
8	felible	G	24	achiral	G	40	tiny	GA
9	inflexible	P	25	2 chiral centers	IT	41	very small	GASC
10	alpha imino	P	26	ionizble	RDCEHKY	42	medium small	VTNDP
11	hydroxyl	STY	27	charged	RDEHK	43	small	GASCVTNDP
12	hydroxyl straight chain	ST	28	acidic	DE	44	large	KRFYW
13	phenol	Y	29	basic	RKH	45	long	KREQ
14	sulfur	CM	30	strong basic	RK	46	very long	KR
15	sulfhydryl	C	31	weak hydrophobilic	STWY	47	medium-long	EQ
16	amide	NQ	32	very hydrophobilic	RNDEQHK	48	short	GASCT

We have addressed the discrimination problem, where given the sequences of a mesophilic protein and a thermophilic or hyper-thermophilic counterpart, the objective is to determine which is which. This was done by assembling a large set of thermophilic protein chains from the PDB and their corresponding mesophilic analogs and another large non-redundant set of hyper-thermophilic PDB protein chains along with their mesophilic analogs. They will be referred to as the pair sets: pairs thermophile and pairs hyper-thermophile. We have computed several sequence based numerical indices, based on the quantities that

other authors have reported that are associated with thermostability. We tested their ability to successfully discriminate between thermophile/mesophile pairs [2].

III. FEATURE VECTOR CREATION

The feature vector of this study is extracted from protein sequence and consists of 468 entries. Twenty features represent frequency of 20 amino acids and are calculated according to equation (1):

$$AA_i = \frac{N_i}{n}, \quad i = 1,2,\dots,20 \tag{1}$$

where N_i is the frequency of *i*th amino acid and n is the length of protein sequence.

Since having twenty different amino acids, there is a maximum of 400 different dipeptide compositions findable in protein sequences [16], and therefore 400 features are assigned to dipeptides. Each feature of this type is calculated according to equation (2):

$$Dep_i = \frac{M_i}{n-1}, \quad i = 1,2,\dots,400 \tag{2}$$

where M_i is the frequency of *i*th dipeptide composition.

Forty-eight features represent chemical-physical attributes of amino acids that for a particular protein, each attribute is obtained by calculating frequency of amino acids relevant to that attribute in the protein. Table (1) shows chemical-physical features used in this study [17].

IV. METHOD

In our proposed method, for the creation of training and testing datasets, k-fold cross validation method has been used [18]. In each iteration of k-fold cross validation, training samples are divided into thermophile, hyper-thermophile and mesophile classes based on their class label.

For each feature, *f*, average of values of that feature in all samples of each class is calculated separately; i.e. we calculate ($\overline{\text{comp}_{f,H}}$) for hyper-thermophile class, ($\overline{\text{comp}_{f,T}}$) for thermophile class and ($\overline{\text{comp}_{f,M}}$) for mesophile class. This procedure is repeated for all features. Therefore, we have a matrix with dimension of 3×468 that its i,j^{th} entry is the average of j^{th} feature for all the samples of *i*th class.

To test the proposed model, at the end of each iteration and for each test sample, feature vector of the sample is created (comp_f). Then, subtraction of values of each feature of test sample and corresponding average of that feature for three classes is calculated. After that, sum of absolute differences for all the features is computed based on equation (3) for mesophile class as well as equation (4) for thermophile class and equation (5) for hyper-thermophile class [19].

$$\sigma_M = \sum_f \left| comp_f - \overline{comp_{f,M}} \right| \tag{3}$$

$$\sigma_T = \sum_f \left| comp_f - \overline{comp_{f,T}} \right| \tag{4}$$

$$\sigma_H = \sum_f \left| comp_f - \overline{comp_{f,H}} \right| \tag{5}$$

According to previous equations, protein is mesophilic if

$$\sigma_M < \sigma_H \text{ and } \sigma_M < \sigma_T \tag{6}$$

and is thermophilic if

$$\sigma_T < \sigma_H \text{ and } \sigma_T < \sigma_M \tag{7}$$

otherwise, we have a hyper-thermophilic protein.

V. EXPERIMENTAL RESULTS

The final performance of proposed method is measured by three evaluation measures: sensitivity, specificity and accuracy. These indices are calculated using equations 8 to 10 [20].

$$sensitivity = \frac{TP}{TP + FN} \tag{8}$$

$$specificity = \frac{TN}{TN + FP} \tag{9}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

As we investigated three classes in this study, sensitivity and specificity are calculated separately for each class. For example, for thermophile class, TP is the number of thermophile samples that our proposed system identifies them correctly as thermophilic proteins. FP is the number of mesophile or hyper-thermophile samples that system identifies them wrongly as thermophilic proteins. TN is the number of mesophile or hyper-thermophile samples that system identifies them correctly as mesophilic or hyper-thermophilic proteins. FN is the number of thermophile samples that system identifies them wrongly as mesophilic or hyper-thermophilic proteins. TP, FP, TN and FN are defined similarly for mesophile and hyper-thermophile classes. The other measure, accuracy, shows the number of sample which system identifies their class correctly.

We applied our proposed method on dataset with four different feature vectors. In the first mode, the feature vector only contained amino acid frequencies.

In second mode, dipeptide composition frequencies; in third mode, physical-chemical features of table (2) and in forth mode, all the previous features were used to build feature vector of samples. We applied this division to study the effect of different bunches of features on discrimination power and evaluation measures. As mentioned earlier, K-fold cross validation was used to create train and test datasets and the results shown in tables are average of different iterations of the method.

Because some applications work only with mesophile and thermophile classes, performance of system in 2-class mode evaluated that the results are shown in tables (3) to (5).

For a more comprehensive review, we have defined a new performance measure called "$SESPAVG_{ij}$" which is arithmetic mean of all sensitivities and specificities of i and j classes for each feature set. i.e., when we are studying the results of applying two-class mode of algorithm on thermophilic and mesophilic class, $SESPAVG_{MT}$ is the arithmetic mean of SE_M, SE_T, SP_M and SP_T. These values have been added to tables (2) to (5) as last column.

Table 2: Results of the proposed method on Three-class mode

	Accuracy	SE_M	SE_T	SE_H	SP_M	SP_T	SP_H	$SESPAVG_{HMT}$
20 Amino Acids	**63.57%**	**61.69%**	**54.18%**	82.26%	**52.87%**	**88.82%**	95.98%	**72.63%**
48 Chemical-Physical	57.62%	59.27%	38.44%	73.88%	50.94%	85.02%	93.94%	66.92%
400 Dipeptide Composition	59.74%	55.92%	50.58%	**84.77%**	50.09%	88.19%	**96.39%**	70.99%
468 All Features	59.05%	59.00%	41.21%	80.37%	51.97%	85.94%	95.32%	68.97%

Table 3: Results of Two-class mode on hyper-thermophilic and mesophilic classes

	Accuracy	SE_M	SE_H	SP_M	SP_H	$SESPAVG_{HM}$
20 Amino Acids	80.83%	78.42%	90.36%	51.19%	97.27%	79.31%
48 Chemical-Physical	**84.00%**	**83.42%**	85.96%	**56.16%**	96.03%	**80.39%**
400 Dipeptide Composition	58.95%	50.02%	**94.89%**	31.83%	97.33%	68.52%
468 All Features	73.68%	68.96%	93.28%	42.90%	**97.63%**	75.69%

Table 4: Results of Two-class mode on thermophilic and mesophilic classes

	Accuracy	SE_M	SE_T	SP_M	SP_T	$SESPAVG_{MT}$
20 Amino Acids	**73.86%**	**76.37%**	64.78%	**44.55%**	88.21%	**68.48%**
48 Chemical-Physical	67.21%	68.75%	62.16%	35.83%	86.54%	63.32%
400 Dipeptide Composition	65.72%	64.09%	**71.45%**	36.81%	**88.47%**	65.21%
468 All Features	72.28%	75.01%	63.25%	42.56%	87.58%	67.10%

Table 5: Results of Two-class mode on hyper-thermophilic and thermophilic classes

	Accuracy	SE_T	SE_H	SP_T	SP_H	$SESPAVG_{HT}$
20 Amino Acids	**83.44%**	**82.66%**	84.03%	**80.36%**	86.30%	**83.34%**
48 Chemical-Physical	66.94%	65.58%	67.85%	62.36%	70.70%	66.62%
400 Dipeptide Composition	77.86%	67.03%	**92.27%**	70.79%	**89.79%**	79.97%
468 All Features	75.90%	71.02%	81.26%	70.95%	81.78%	76.25%

VI. CONCLUSION

Studying the frequency of amino acids in hyper-thermophilic, thermophilic and mesophilic proteins shows that Gly, Ser and Thr are more frequent in mesophilic proteins. Arg, Leu and Pro amino acids are more frequent in thermophilic proteins and Glu, Iso and Lys amino acids are more frequent in hyper-thermophilic proteins.

The other result of this study is that among 400 dipeptide amino acid compositions, AA, LL, LA, AL, QA, QL, AQ, LT, TL and EQ are more frequent in mesophilic proteins; IE, EE, EK, KE, VE, EI, KI, KK and VK are more frequent in thermophilic proteins and RI, AK, LA, YI, YE, RK and SK are more frequent in hyper-thermophilic proteins. In addition, Amino acids with charge and acidic attributes are more frequent in hyper-thermophilic proteins.

Studying SESPAVG measures and according to the results of table (2) which is related to three-class mode, values of Accuracy and $SESPAVG_{HMT}$ is maximized if we use 20 amino acid frequency features to discriminate samples. In two-class mode, in table (3), accuracy and $SESPAVG_{HM}$ are maximized if we use physical-chemical features and in tables (4) and (5), accuracy, $SESPAVG_{MT}$ and $SESPAVG_{HT}$ are maximized if we use 20 amino acid frequency features.

So this can be concluded that 20 amino acid frequency features are better that other features for discriminating thermophile class samples from mesophile class samples and hyper-thermophile class samples from thermophile class samples. Also, physical-chemical features are better than other features for discriminating hyper-thermophile class samples from mesophile class samples.

Because the number of entries of 20 amino acid frequency and physical-chemical feature vectors which are suggested based on results are lower than 400 dipeptide compositions and 468 all feature vectors, required time for identifying the class of a new protein using former vectors is also less than latter vectors.

Dataset used in this research include three-dimensional structure of proteins in addition to convention amino acid sequence. This advantage makes possible the use of features extracted from the 3D structure in future researches.

VII. REFERENCES

[1] C-I Branden, J.Tooze, "Introduction to protein structure", 2nd edition, New York: Garland Pub., 1999.

[2] Todd T, Iosif V, "Discrimination of thermophilic and mesophilic protein", Taylor and Vaisman BMC Structural Biology, vol. 10, 2010.

[3] Xingyu, W., Shouliang, C., Mingde, G., "General Biology", Version 2, Higher Education Press, Beijing, 2005.

[4] Brock T, Freeze H, "Thermus aquaticus gen. n.and sp. n.., a Nonsporulating Extreme Thermophile", J Bactriol, vol. 98, pp. 289-297, 1969.

[5] Jingru Xu, Yuehui Chen, "Discrimination of Protein Thermostability Based on a New Integrated Neural Network", vol. 1, pp. 107-112, Springer, 2011.

[6] Kumar S, Nussinov R, "How do thermophilic proteins deal with heat?", Cell Mol Life Sci, vol. 58, pp. 1216-1233, 2001.

[7] Thompson MJ, Eisenberg D, "Transprotemic evidence of a loop-deletion mechanism for enhancing protein thermostability", J Mol Biol, vol. 290, pp. 595-604, 1999.

[8] Haney PJ, Jonathan HB, Berald LB, "Thermal adaption analyzed by comparison of protein sequences from mesophilic and extremely thermophilic Methanococcus species", PNAS, vol. 96, pp. 3578-3583, 1999.

[9] Vielle C, Zeikus GJ, "Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability", Microbiol Mol Biol Rev, vol. 65, pp. 1-43, 2001.

[10] Ding YR, Cai YJ, Zhang GX, "The influence of dipeptide composition on protein thermostability", FEBS Lett, vol. 569, pp. 284-288, 2004.

[11] Kumarevel TS, Gromiha MM, Ponnuswamy MN, "Structural class predication: an application of residue distribution along the sequence", Biophys Chemist, vol. 88, pp. 81-101, 2000.

[12] Gromiha MM, Shandar A, Makiko S, "Application of residue distribution along the sequence for discriminating outer membrane proteins", Comput Biol Chem, vol. 29, pp. 135-142, 2005.

[13] Seung P, Young J, "Protein Thermostability: Structure-Based Difference of Amino acid between Thermophilic and Mesophilic Proteins", Elsevier Journal of Biotech, vol. 111, pp. 269-277, 2004.

[14] Michael G, Xavier S, "Discrimination and Classification of Mesophilic and Thermophilic Protein using Machine Learning Algorithms", Willy InterSience, pp. 1274-1279, 2007.

[15] http://www.pdb.org. Available at June 24, 2012.

[16] Zahng G, Fang B, "Study on the Discrimination of Thermophilic and Mesophilic Proteins Based on Dipeptide Composition", Chinese journal of biotechnology, vol. 22, pp. 293-298, 2006.

[17] Eshkin A, Ghafuri H, "Predication of relative solvent accessibility by support vector regression and best-first method", Excli Journal, vol. 9 , pp. 29-38, 2010.

[18] Platzer A, Percpo P, "Characterization of protein-interaction networks in tumors", BMC Bioinformatics, vol. 8, 2007.

[19] Szilagyi A, Zavodszky P, "Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of comprehensive survey", Elsevier, vol. 8, pp. 493-504, 2000.

[20] Chen Yu, Han, Kyungsook, "BSFINDER: Finding Binding Sites of HCV Proteins Using a Support Vector Machine", Protein and Peptide Letters, vol. 16, pp. 373-382(10), 2009.